An Approach for Accessing Linked Open Data for Data Mining Purposes

Andreas Nolle, German Nemirovski
Albstadt-Sigmaringen University
nolle, nemirovskij@hs-albsig.de

Abstract

In the recent time the amount of semantic data publicly available online grows continuously. This data is a valuable source of information that can be inferred by means of different technologies. Apart of reasoning on data semantics classical OLAP or probabilistic data mining techniques can be applied for pattern recognition, prediction or clustering. However the conversion of linked (open) data basically stored as RDF triples to data structures appropriate for data mining, such as propositional representations or flat tables, appears to be a challenging task. In this paper we propose an approach addressing this task. The approach has been implemented in form of an operator for the RapidMiner software, a data mining environment widely used in the industry and research.

1 Introduction

The idea to close the gap between two technologies for data analysis, like ontology based inference on the one hand and data mining on the other hand, wins interest of scientists increasingly. While the statistical methods comprised under the umbrella term “data mining” are known since decades, the birth of so called semantic technologies embracing ontology-based knowledge representation and inference is connected with the paper of T. Gruber [1]. This paper was published in 1995 and introduces the term *ontology* in the knowledge representation context related to engineering sciences.

One year later Fayyad et al. [2] brought both terms together. However, ontologies are only mentioned in the concluding part of this paper as a perspective technology “allowing the use of prior human knowledge about the underlying

process by the KDD[1] system“. One of the first attempts to integrate ontology-based knowledge representation into the data mining process has been presented two years later by Simeon and Maher [3]. In their application case an ontology provided “context, structure and relationships for representation and integration of discovered patterns” identified in multimedia data related do the domain of building design.

In the recent time the interdisciplinary research on the edge of data mining and ontology-based inference has been progressing significantly. One of the reasons for this progress is doubtless substantial growth of the Linked Open Data cloud [4]. This data is basically stored in RDF format and accessible over numerous SPARQL end points. In this connection a challenging task is to query the RDF data and to convert it to structures and formats which can be processed by data mining tools. The approach described in this paper is addressing this task. In particular it targets the retrieval and integration of RDF data into the processes designed using RapidMiner, a data mining environment widely used in the industry and research.

The paper is structured as following: In the section 2 it gives an overview of approaches related to the proposed one. Section 3 contains the technical description of the work done. Finally the section 4 describes conclusions and new challenges related to the preprocessing of RDF data for data mining purposes.

2 Related Work

In spite of the high interest of the scientific community and strong demand in various application domains, currently there exist not many practically applicable approaches targeting querying of RDF data sources and transforming query answers into formats required by data mining processes, e.g. by RapidMiner processes.

One of these approaches, RMonto [5, 6, 7], is an extension for RapidMiner. RMonto facilitates accessing semantic data and its further processing using machine learning algorithms. Since the data isn’t transformed into the internal data format of RapidMiner this approach is more similar to semantic data mining. Due to differences in data formats numerous data mining operators available in RapidMiner cannot be combined with RMonto. Therefore the existing operators have to be adapted or new operators have to be developed in order to facilitate sophisticated data mining processes basing on RMonto, which is a significant limitation of the approach.

Another approach called rapidminer-semweb has been developed by Khan et al. [8, 9]. The implementation in form of a RapidMiner operator is available online.

[1] KDD stands for Knowledge Discovery in Databases that is a comprehensive process comprising apart of data mining, selection and transformation of data as well as interpretation of analysis results. Though, in the recent time terms KDD and data mining in colloquial language are often used as synonyms.

Though, this approach is strongly related to the task of RDF data retrieval, it is, unfortunately, not mature enough to be applied in practice. Our evaluation of this operator has shown that both methods i) RDF file access, i.e. loading of RDF files from a local directory as well as ii) sending requests to SPARQL endpoints ended up with an error.

The third approach implemented by Paulheim & Fürnkranz [10] rather focuses on enrichment of statistical data with ontology based "semantic" metadata available online, e.g. within the Linked Open Data cloud. However such metadata isn't used as immediate input for data mining processes. Therefore this approach doesn't address the task of RDF data retrieval directly.

Since this survey of available technologies did not deliver any satisfying results the demand on development of software making the RDF data available for a data mining processes in particular for processes developed in the RapidMiner environment is evident.

3 Technical Description

The RapidMiner operator proposed in this paper has been developed to query RDF data sources accessible over SPARQL endpoints and to transform the query answers into proprietary RapidMiner format "understandable" for all other operators available in the RapidMiner environment.

The operator has three parameters. The first one specifies the URL of the SPARQL endpoint that has to be queried. The second parameter contains the SPARQL query itself (Figure below) and the third one is to specify variables of the SPARQL query that have to be interpreted as polynominals.

Fig. 1. Input parameters for the RDF retrieval operator.

To query more than one SPARQL endpoints a federation engine like ELITE [11] can be used. Another option for querying multiple data sources simultaneously is to use SPARQL 1.1's federated queries[2]. Since in practice, formulation of SPARQL queries is a sophisticated process and most of data mining experts aren't expected to have sufficient experience in this field, we have developed an assistance tool that helps the users in query generation issues. Using this tool queries can be generated by selecting ontology's concepts, object and data properties in interaction.

[2] http://www.w3.org/TR/sparql11-federated-query/

After a query has been evaluated by an SPARQL endpoint or by a federation engine and the evaluation result has been retrieved, it is transformed into the internal representation of RapidMiner. The data type of each result variable is analyzed and transformed into the corresponding RapidMiner type. For instance Boolean variables are transformed into binominals and URIs of individuals are transformed into strings. Since polynominal variables can't be identified on the basis of the RDF data type analysis, additional information is required to solve this issue. To identify polynominal variables we use the semantics of data, i.e. particular conceptualization aspects expressed over the ontology's TBox. For this purpose, we have defined a special ontology concept that subsumes all concepts whose individual's URIs should be converted to polynomial variables. Identified polynominals are delivered to the operator using the appropriated operator parameter.

Having the data transformed into the internal representation of RapidMiner, the data can be further processed with conventional operators. To avoid recurrent execution of SPARQL queries, which may be time consuming, we store the query results in a database (Figure 2). Doing so, the data retrieved from the RDF sources previously can be simply reloaded and processed.

Fig. 2. Storing the query results in a database.

As a rule, the SetRole operator should be set immediately after the RDF data is loaded from the database to define specific roles, e.g. *label* or *id* for the particular data.

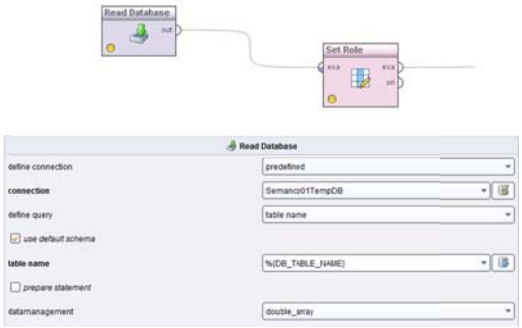

Fig. 3. SetRole Operator following the data reloading operator.

To set the roles we have developed a tool that analyzes the semantics of variables and automatically deduces the appropriate role (Figure 4).

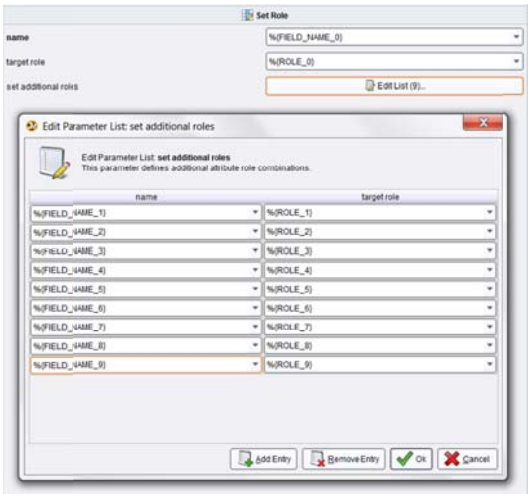

Fig. 3. Defining parameters for the determining of roles of variables

4 Conclusion and Further Work

In this paper we have described a RapidMiner operator for retrieval of RDF data distributed in sources exposing a SPARQL endpoint. After retrieving data in RDF format the operator transforms it into the internal RapidMiner format and so facilitates processing of this data by means of other operators available in the RapidMiner environment.

The implementation has been successful and the operator is now applied for retrieval of data related to the scope of SEMANCO project[3] focusing on carbon reduction in urban planning.

In our future work we purpose to integrate the mentioned tools for query generation and ontology analyses in the RapidMiner operator described in this paper.

Furthermore, apart of probabilistic data mining techniques we plan to facilitate data processing using OLAP technologies. To do so issues related to data mart generations need to be addressed, e.g. through exploiting formally specified data semantics for the automation of data cube design, in particular tackling such challenges as selection and combination of data cube dimensions or summarizability problem as already shown by Romero and Abelló [12].

[3] http://www.semanco-project.eu

Acknowledgements

The main contribution of this work has been developed within the SEMANCO project, which is being carried out with the support of the Seventh Framework Programme “ICT for Energy Systems” 2011–2014, under the grant agreement no. 287534.

References

[1] T. Gruber, „Towards principles for the design of ontologies used for knowledge sharing.,“ *International Journal of Human Computer Studies, Vol. 43 (5-6)*, pp. 907-928, 1995.

[2] U. P.-S. G. S. P. Fayyad, „From data mining to knowledge discovery: An overview.,“ in *U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (eds)*, Bostin, MA, Advances in Knowledge Discovery and Data Mining AAAI Press, 1996, pp. 1-34..

[3] S. J. a. M. M. L. Simoff, „Ontology-based multimedia data mining for design information retrieval.,“ *Proc. Computing in Civil Engineering, ASCE, Reston*, p. 212–223, 1998.

[4] [Online]. Available: http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData. [Zugriff am 14 June 2013].

[5] J. Potoniec und A. Lawrynowicz, „RMonto: ontological extension to RapidMiner,“ in *Demo session at the International Semantic Web Conference*, 2011.

[6] J. Potoniec und A. Lawrynowicz, „RMonto-towards KDD workflows for ontology-based data mining,“ in *eCML PKDD 2011*, 2011.

[7] [Online]. Available: http://semantic.cs.put.poznan.pl/RMonto/doku.php.

[8] M. A. Khan, G. A. Grimnes und A. Dengel, „Two pre-processing operators for improved learning from semanticweb data,“ in *First RapidMiner Community Meeting And Conference (RCOMM 2010)*, p. 2010.

[9] [Online]. Available: http://code.google.com/p/rapidminer-semweb/.

[10] H. Paulheim und J. Fürnkranz, „Unsupervised generation of data mining features from linked open data,“ in *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, ACM, 2012.

[11] A. N. G. Nolle, „ELITE: An Entailment-based Federated Query Engine for Complete and Transparent Semantic Data Integration.,“ *Proc. of International Workshop on Description Logic.*, July 2013.

[12] O. R. a. A. Abelló, „Generating Multidimensional Schemas from the Semantic Web,“ 2007.